# TARRANT COUNTY EVALUATION STUDY – FINAL REPORT

DAVID BUCKERIDGE, AMAN VERMA, AND DAVID SIEGRIST

Abstract. The study described in this report is one component of a larger effort to evaluate the effectiveness of a regional syndromic surveillance reporting network in North Central Texas that is currently housed in the offices of the Tarrant County Public Health Department (Fort Worth, Texas). In the current study, we aimed to assess the ability of the surveillance network to detect an inhalational anthrax disease outbreak through temporal surveillance.

To evaluate outbreak detection, we used the 'inject' approach. This entailed generating many simulated disease outbreaks, superimposing each simulated outbreak onto real background data from the surveillance network, and then feeding the combination of real and simulated data to a model of the syndromic surveillance network. We examined the ability of the surveillance network at different alarm rates to detect disease outbreaks that would infect 100, 500, 1,000, or 5,000 people.

The results suggest that the surveillance network is capable of detecting a large proportion of inhalational anthrax outbreaks infecting as few as 100 people at false alarm rates acceptable to the surveillance team. For outbreaks of all sizes examined, the surveillance network saved a considerable proportion of time, or detected the outbreak faster than clinical case-finding, for many of the simulated outbreaks. This ability of the surveillance network to provide an early indication of an outbreak was particularly pronounced for smaller outbreaks where clinical case-finding tended to take longer to identify the sentinel case, and when the surveillance network was operating at a higher alarm rate.

## 1. Introduction

1.1. **Study Motivation.** The study described in this report is one component of a larger effort to evaluate the effectiveness of a regional syndromic surveillance reporting network in North Central Texas that is currently housed in the offices of the Tarrant County Public Health Department (Fort Worth, Texas). The focus of the larger evaluation effort is to assess the performance of the syndromic surveillance network as it currently operates, to identify strengths and weaknesses of the network, and to suggest modifications to the network that may improve performance. In the study described in this report, we aim to assess the ability of the current network to detect an inhalational anthrax disease outbreak through temporal surveillance. In the remainder of this introductory section, we describe the surveillance network (Section 1.2) and provide an overview of the contents of this report (Section 1.3).

1.2. **Overview of North Central Texas Syndromic Surveillance Network.** The North Central Texas Syndromic Surveillance Network had 30 hospitals reporting ED visits during the study interval from July 5th, 2004 - March 8th 2006.

*Date*: August 14, 2006.

The majority of these hospitals (16, 53%) were reporting each visit in real time using the HL7 standard over the Internet, while the remaining hospitals (14, 47%) were reporting visits once daily by a batch file sent at midnight. The 30 hospitals that contributed data to the regional surveillance network at the time of the study are spread across a 200-square-mile region of North Central Texas. Most of the reporting hospitals are located in the 16-county region that composes the North Central Texas Council of Governments (NCTCOG) and encompasses the cities of Dallas and Fort Worth. That region's current population is 6,242,800 (NCTCOG estimate).

The 30 reporting hospitals, though representing about 45 percent of the total number of hospitals with EDs in the region (based on a list published by the Dallas-Fort Worth Hospital Council), include most of the area's largest, busiest facilities. The reporting hospitals handled more than 1.1 million emergency department visits in 2005, according to published reports on their Websites. (That total includes 465,000 at 12 Texas Health Resources facilities, 261,000 at eight Baylor Health Care System facilities, 116,000 at three Cook Children's Health Care System facilities, 115,000 at two Methodist Health System facilities, 100,000 at two Health Management Associates facilities, 80,000 at John Peter Smith (JPS) Hospital, 25,000 at USMD Hospital in Arlington, and 21,000 at UT Southwestern Medical Center.)

The surveillance network analyzes the data routinely 10 times each day, and in practice generates two or three alerts each week. Every analysis occurs on three levels, examining fluctuations in case counts for each ZIP code, each county, and the entire region. So, although it would be unusual, it is possible for a large spike in case counts for a single ZIP code to generate up to three alerts (one for the ZIP code, another for the county, and a third for the region). The four largest cities in the region alone account for more than 200 distinct ZIP codes (121 in Dallas, 58 in Fort Worth, 19 in Arlington, and 9 in Plano).

The 38 users of the regional surveillance system work at seven different public health agencies located throughout North Central Texas, one of which is a regional office of the Texas Department of State Health Services. The system users are mostly epidemiologists (74 percent), but the group also includes two health authorities, a medical director, a public health department director, a health educator, a Geographic Information Systems (GIS) administrator, and three staff members in the Southwest Center for Advanced Public Health Practice (the grant-funded unit of Tarrant County Public Health that's responsible for developing and maintaining the surveillance system).

Over the study interval, the average daily number of records reported by each hospital ranged from 14 to 216 and the average daily number of records reported by all hospitals combined was 3,992. The surveillance network uses several applications, but the only one tested for this report is the Real-Time Outbreak Disease Surveillance (RODS) software, a system supported with funding from the Department of Homeland Security and developed by the University of Pittsburgh and Carnegie Mellon University. [18]. The RODS software is used for data collection, classification of visit records into syndromes, and statistical analysis of aggregated records to detect disease outbreaks. When records are received, the RODS system classifies the free-text chief complaint in each record into a syndrome using the CoCo chief complaint classifier. The RODS system then analyzes the aggregated data 10 times each day (midnight, 6 am, 8 am, 10 am, noon, 2 pm, 4 pm, 6 pm, 8

pm, and 10 pm) using two temporal algorithms,[1] a cumulative sum and an recursive least-squares (RLS) algorithm. The specificity of each algorithm is controlled by altering a threshold parameter.

When a statistical aberration is signaled, epidemiologists follow a response protocol that includes a brief initial investigation, followed by a more detailed investigation only if warranted. The initial investigation involves looking at cases to determine if they cluster in space, examination of a line listing of the records contributing to the alert, and looking at the pattern of alerts across syndromes.

1.3. **Report Overview.** In the next section of this report, we describe briefly the methodology (Section 2) and readers are directed to the Appendix for a detailed description (Section 7). In Section 3, we present the main findings from the report, including the performance of the surveillance network by the alarm rate, the size of the disease outbreak, and other surveillance network characteristics. Finally, in Section 4, we conclude with recommendations for modifying the current network to improve performance and by identifying areas for future evaluation.

## 2. Study Overview

2.1. **Study Design.** To evaluate outbreak detection, we used the 'inject' approach. This entailed generating many simulated disease outbreaks, superimposing each simulated outbreak onto real background data from the surveillance network, and then feeding the combination of real and simulated data to a model of the syndromic surveillance network. We relied on simulated outbreak data, superimposed onto real data, for the evaluation because data from real outbreaks, especially those due to inhalational anthrax, are available in neither the form nor the quantity needed for a rigorous evaluation.

The simulated outbreaks reflected the additional emergency department (ED) visits that we would expect to occur following an aerosol release of anthrax spores. To simulate an outbreak, we set the number of individuals infected and then simulated disease progression and ED visits for each infected person. A simulated ED visit included a time and a syndrome assigned to the visit. The baseline data were records of ED visits obtained from the surveillance network. We used the CoCo chief-complaint classifier from RODS to categorize the baseline records into syndromes and statistical algorithms from RODS to analyze time-series of aggregated baseline and simulated records [18]. In addition to assessing outbreak detection through syndromic surveillance, we also simulated the time to clinical diagnosis of the first case in each outbreak and used these data to compare syndromic surveillance to clinical case-finding.

In the study, we examined the ability of the surveillance network at different alarm rates to detect disease outbreaks that would infect 100, 500, 1,000, or 5,000 people. Before presenting the results, we provide some context for interpreting the alarm rate (Section 2.2) and the number infected (Section 2.3).

2.2. **Interpreting the Alarm Rate.** The alarm rate is the frequency at which the surveillance network alarms in the absence of a disease outbreak; this can also be expressed as the background alarm rate. In general, higher alarm rates result in better detection performance (i.e., higher sensitivity and faster time to

[1] A *temporal algorithm* is used to analyze counts or rates recorded at consecutive points in time. Information on the spatial location of the data is not used by a temporal algorithm.

detection). The alarm rate can be changed by adjusting the threshold parameter of the statistical method that is used to detect outbreaks. In the statistical literature, the alarm rate is defined as the complement of specificity, which is defined as the probability of not sounding an alarm when there is in reality no outbreak occurring. The number of alarms per a unit of time, such as a week, month, or year is a more intuitive measure of the alarm rate than specificity.

There are at least four important characteristics of the alarm rate that should be kept in mind when interpreting detection performance. First, an alarm indicates a statistical aberration in the data and not necessarily a true disease outbreak. Alarms that occur in the absence of any known disease outbreak may be due to unexpected changes in data reporting from hospitals, natural variation in the data, or true disease outbreaks that are not recognized by public health. Second, we assumed for this study, that all alarms in the baseline data (i.e., alarms not occurring during a simulated outbreak) were false alarms. This is a reasonable assumption if one is interested in detecting only outbreaks due to bioterrorism, but it may lead to conservative estimates of specificity if one is interested in using the network to detect other types of outbreaks as well. Third, different alarm rates may be acceptable in different surveillance settings. For example, in a local or regional surveillance network, such as the North Central Texas network, higher alarm rates may be acceptable if epidemiologists can rule-out alarms with a minimal amount of effort. Finally, while specificity is not affected greatly by the frequency of analysis, the alarm rate per unit of time is influenced strongly by the frequency of analysis. For example, a specificity of 0.9 translates into 7 alarms per week when the data are analyzed ten times each day and less than 1 alarm per week when the data are analyzed once each day.

2.3. **Interpreting the Number Infected.** The number infected is the total number of people that are infected and develop symptoms following a simulated exposure. In general, only a proportion of infected individuals will visit an emergency department during the course of their illness. Moreover, these visits will occur over time and symptomatic individuals will tend to visit different hospitals. Finally, when individuals do visit a hospital, they will likely be assigned a variety of diagnoses, especially in the early stages of disease, when symptoms of inhalational anthrax are non-specific.

The result is that the number of visits 'seen' by the surveillance network is smaller than the number infected and those visits are spread out over many days. For example, in a simulated outbreak where 1,000 people were infected, these simulated individuals made approximately 880 visits to emergency departments with 270 visits before the peak of the epidemic curve, which occurred, on average, 10 days following exposure. Of the visits that occurred before the peak of the epidemic curve, approximately 192 were coded as being for a respiratory condition. So, the average number of additional respiratory visits that were 'seen' by the surveillance network was approximately 19 each day, superimposed on an average baseline incidence of 413 respiratory visits each day for an additional 'signal' of 4.6% over the baseline.

A reasonable question is when the increase in incidence due to an outbreak would be noticed by emergency department staff (note this is a different question than whether a sentinel case would be diagnosed by an astute clinician). Continuing with the numbers used above, if there were an additional 19 cases of respiratory

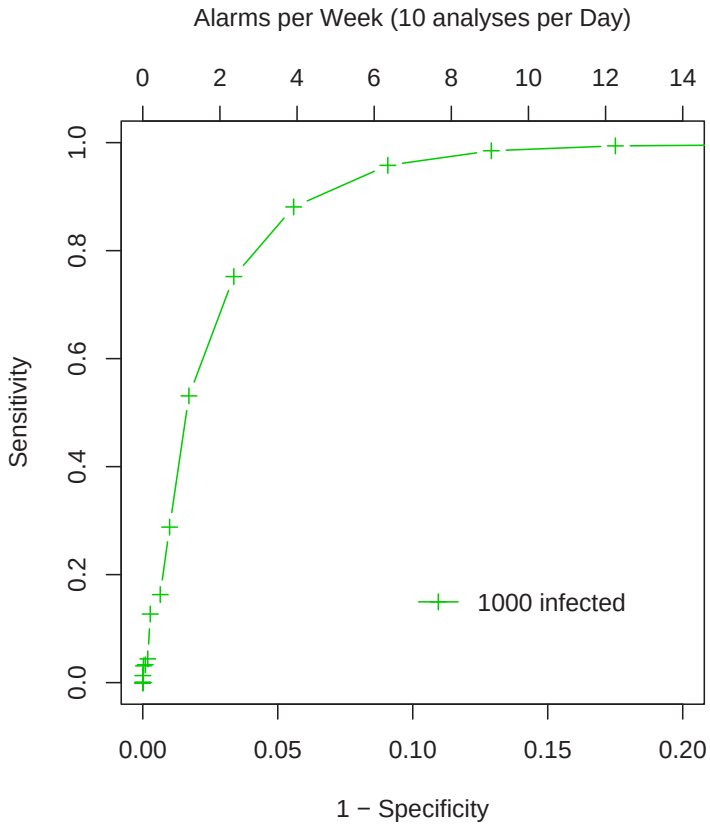


FIGURE 1. The Receiver Operating Characteristics (ROC) curve for outbreak detection with 1,000 people infected. Note that more outbreaks are detected (i.e., the sensitivity improves) as the alarm rate increases.

disease seen on a given day, and these cases were distributed uniformly across the 30 hospitals in the surveillance network, then we would expect each emergency department to 'see' an additional 0.63 cases per day spread across the various staff working in the emergency department.[2] It is difficult to say with certainty whether this additional number of cases would be noted to be unusual, but these numbers make clear that an outbreak infecting 1,000 people is likely to result in a relatively small number of additional cases at any one hospital.

## 3. Main Results

In our analysis, we considered multiple clinical presentations that might occur following an aerosol anthrax release and we also considered two different statistical algorithms that might be used in the surveillance system. We present the results for

[2]In reality, the distribution will not always be uniform. As well, the average of 0.63 cases per hospital can be interpreted as 1 or 0 cases per hospital.

| Number Infected | 1 - Specificity (Alarms per Week) Required to Detect Proportion | | |
|---|---|---|---|
| | **50%** | **80%** | **95%** |
| 100 | 0.023 (1.6) | 0.060 (4.2) | 0.117 (8.2) |
| 500 | 0.022 (1.5) | 0.052 (3.6) | 0.107 (7.5) |
| 1,000 | 0.016 (1.1) | 0.042 (2.9) | 0.087 (6.1) |
| 5,000 | 0.002 (0.2) | 0.007 (0.5) | 0.017 (1.2) |

TABLE 1. The specificity, or alarms per week, required to detect 50%, 80%, and 95% of outbreaks by four different numbers infected. Note that the majority of outbreaks of all sizes examined are detected at reasonable alarm rates, but detection of a large proportion of smaller outbreaks requires a high alarm rate.

only the respiratory syndrome and the statistical algorithm (recursive least-squares or RLS) that had the best detection performance.

3.1. **Outbreak Detection by Alarm Rate.** A greater proportion of outbreaks were detected when more alarms were tolerated. Figure 1, a conditional Receiver Operating Characteristics (ROC) curve, demonstrates this relationship for the situation where 1,000 people were infected. The trade-off between alarm rate and the proportion of outbreaks detected (or sensitivity) is illustrated for selected detection proportions in Table 1. To detect half of the simulated outbreaks, alarm rates of between 1.6 per week (for 100 people infected) and 0.2 per week (for 5,000 people infected) were required. These alarm rates are probably acceptable in most settings, even for the smaller size outbreaks. Higher alarm rates were required, however, to ensure detection of a large proportion of the simulated outbreaks. For an outbreak that infected 100 people, and alarm rate of 8.2 per week (out of 70 analyses per week) was required to detect 95% of outbreaks.[3] With an outbreak that infected 5,000 people, the alarm rate required to detect 95% of outbreaks was still relatively low, 1.2 per week.

3.2. **Outbreak Detection by Number Infected.** As one would expect intuitively, detection performance improved as the number of individuals infected increased. Figure 2 shows the slight improvement in accuracy as the number of people infected increased from 100 to 1,000 and the more marked improvement between 1,000 infected and 5,000 infected. Perfect detection would result in an ROC curve that rises instantly to a sensitivity of 1 at 0 alarms per week. One approach to quantifying the relative performance for different sizes of outbreaks is to calculate the area under the ROC curve for each size of outbreak. Perfect detection would result in an area under the curve (AUC) of 0.2. Table 2 (see page 11) displays the AUC for each line in Figure 2. Accuracy improves, or the AUC increases,

[3]This alarm rate is determined, in part, by the number of analyses performed each day. See Section 3.4 for a discussion of how fewer analyses per day will lower the alarm rate

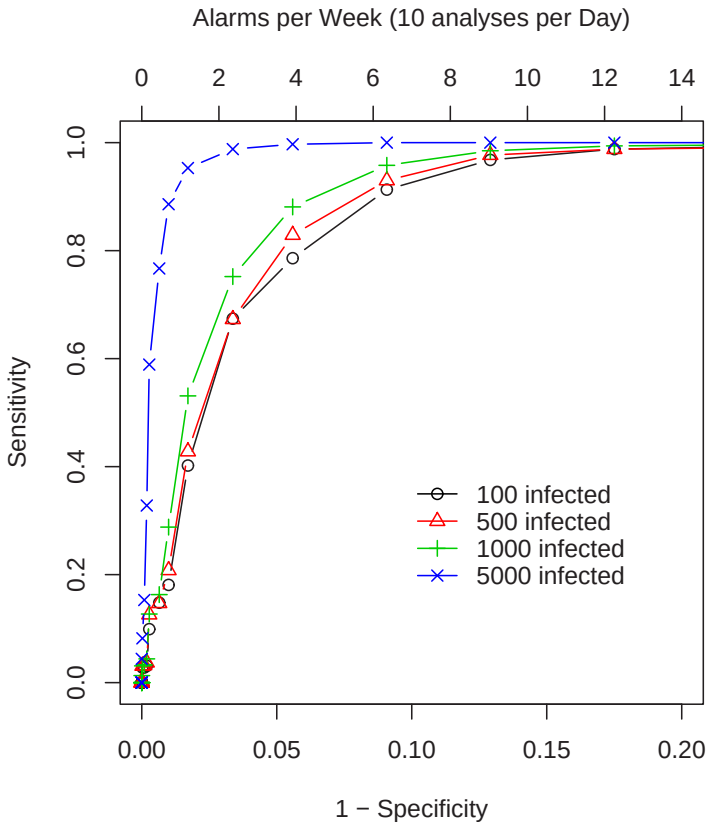


FIGURE 2. The Receiver Operating Characteristics (ROC) curves for outbreak detection with different numbers of people infected. Note that more outbreaks are detected (i.e., the sensitivity improves) at a given alarm rate as the number infected increases.

as the number of people infected increases, and detection performance with the surveillance network is nearly perfect for an outbreak that infects 5,000 people.

Another useful measurement of detection performance is the comparison between outbreak detection through the surveillance network and outbreak detection through clinical case-finding. In other words, comparing the initial alarm from the surveillance network to the first case diagnosed through the routine use of blood culture[4]. As with detection through syndromic surveillance, the simulated time to outbreak detection through clinical case-finding decreased as the number of infected individuals increased. Figure 3 demonstrates this relationship, and the median time until outbreak detection through clinical case-finding decreased from 7.5 days when 100 people were infected to 4.6 days when 5,000 people were infected.

[4]See the Appendix (Section 7) for a description of the model used to simulated clinical case-finding

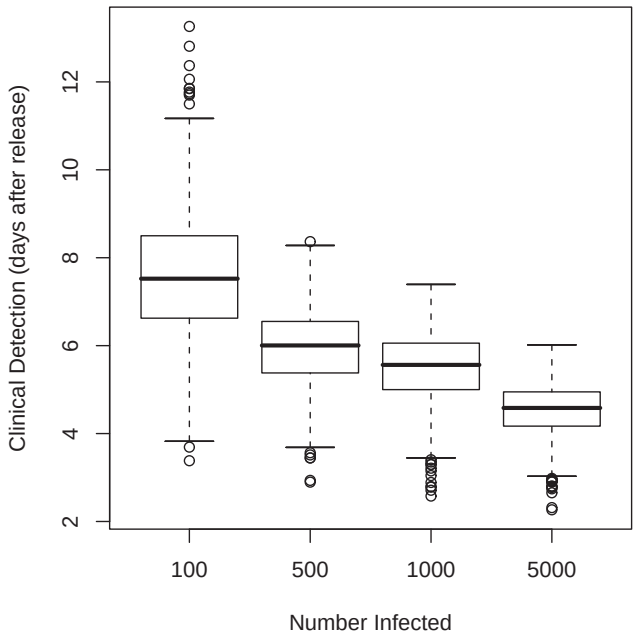


FIGURE 3. Boxplots showing the distribution of the simulated time to clinical detection of outbreaks through routine blood culture testing. The horizontal line is the median value, the box is the interquartile range, the whiskers are the range of the data and the the circles are outlying points. Note that the median time to clinical detection of an outbreak decreases as the number infected increases.

To compare detection through syndromic surveillance and clinical case-finding directly, we plotted the 3-dimensional analogue to the ROC curve, the Timeliness Receiver Operating Characteristics (TROC) surface, and calculated the volume under this surface, which is analogous to the AUC. The third axis in the TROC plot (Figure 4) measures the cumulative proportion of time saved due to outbreak detection through the surveillance network as compared to clinical case-finding. For example, if clinical case-finding detected an outbreak 5 days following release of spores and syndromic surveillance detected the same outbreak 4 days after release, surveillance would save 20% of the time. The TROC plot displays the proportion of outbreaks detected (sensitivity or TP, on the z-axis) with a given proportion of time saved (1 - proportion of time, on the x-axis) over a range of alarm rates (FP, on the y-axis). The top-right corner in Figure 4 demonstrates, for example, that some time was saved (i.e., 1 - Proportion of time saved was $\leq 1.0$) in over 80% of outbreaks when 1 - specificity (or FP) rose to 0.2.

Another perspective on the proportion of time saved is to take 2-dimensional 'slices' through the TROC surface at set alarm rates. Figure 5 demonstrates 3 such

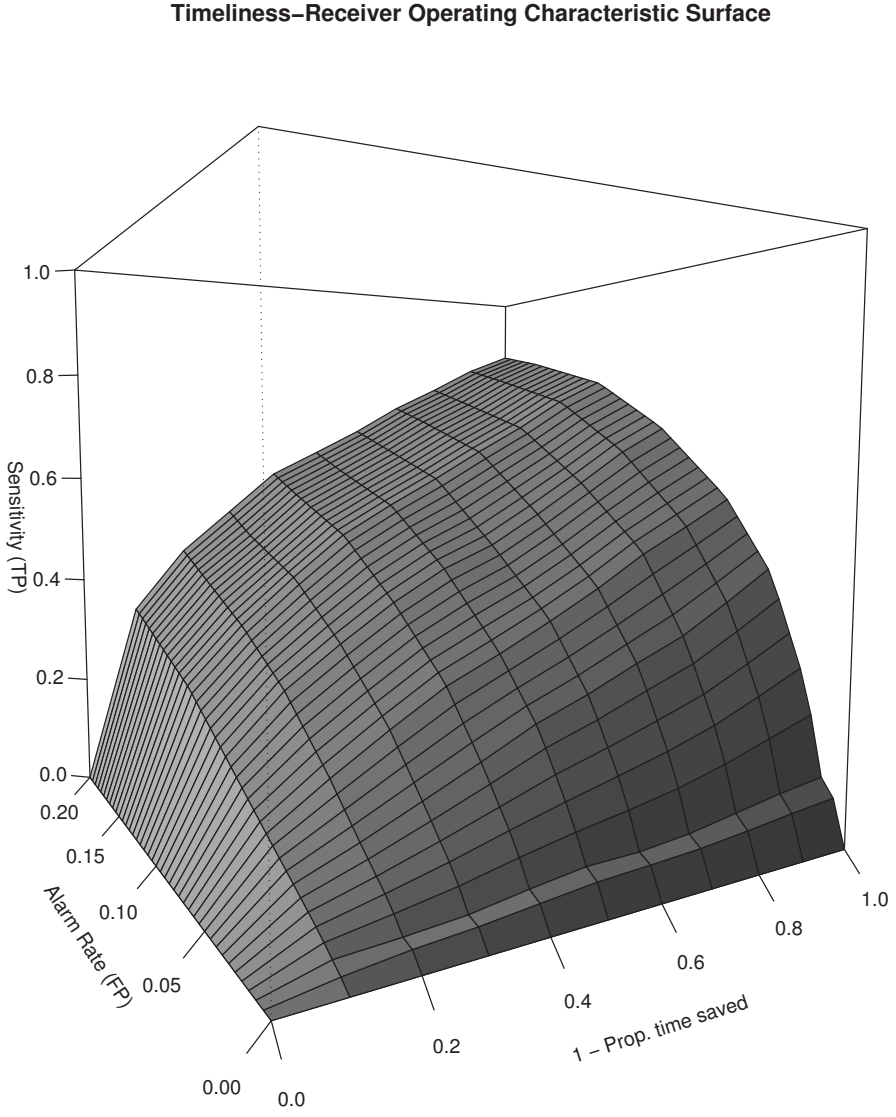


FIGURE 4. The Timeliness Receiver Operating Characteristics Surface for 1,000 people infected. The surface shows, at different alarm rates (FP) the frequency (TP) with which a proportion of time was saved through surveillance as compared to outbreak detection through clinical case-finding. The top-right corner demonstrates, for example, that some time was saved (i.e., 1 - Proportion of time saved was $\leq 1.0$) in over 80% of outbreaks when 1 - specificity (or FP) rose to 0.2. See Figure 5 for 2-dimensional slices through the surface at different alarm rates.

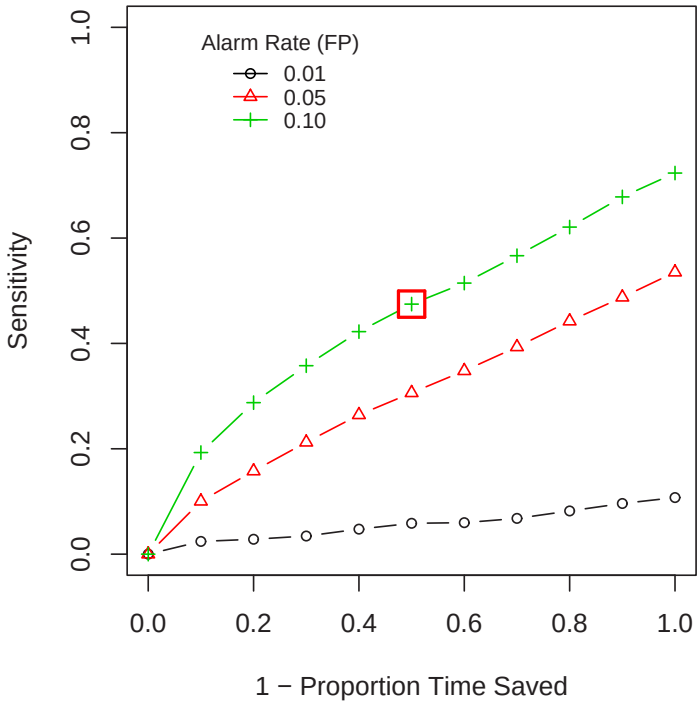


FIGURE 5. 2-Dimensional 'slices' through the the Timeliness Receiver Operating Characteristics Surface shown in Figure 4. These 2-Dimensional slices demonstrate the proportion of outbreaks (Sensitivity) for which a given amount of time was saved through syndromic surveillance relative to clinical case-finding (1 - Proportion of Time Saved) at 3 alarm rates. The point highlighted with the red box indicates that when operating at an alarm rate of 0.10, surveillance saved at least 50% of the time (i.e., 1 - Proportion of Time saved, on the horizontal axis) in nearly half of the simulated outbreaks (i.e., Sensitivity, on the vertical axis).

slices at alarm rates of 0.01, 0.05, and 0.10. From this perspective, it is clear that outbreak detection through syndromic surveillance saved time over clinical case-finding in fewer than 10% of outbreaks when the surveillance system was operating at an alarm rate of 0.01. When the alarm rate was set higher, to 0.10 however, the surveillance system saved more time as compared to clinical case-finding. The point highlighted with the red box in Figure 5 indicates that when operating at an alarm rate of 0.10, surveillance saved at least 50% of the time between outbreak onset

| Infected | Area Under Curve | Accuracy | Volume Under Surface | Timeliness Accuracy |
|---|---|---|---|---|
| 100 | 0.164 | 0.82 | 0.0934 | 0.47 |
| 500 | 0.166 | 0.83 | 0.0831 | 0.42 |
| 1,000 | 0.173 | 0.86 | 0.0794 | 0.40 |
| 5,000 | 0.195 | 0.97 | 0.0709 | 0.35 |

TABLE 2. The area under the Receiver Operating Characteristics (ROC) curve and the volume under the Timeliness Receiver Operating Characteristics Surface. The area, with a maximum of 0.2 in this case, indicates the overall accuracy of the system in detecting outbreaks for each level in the range of number of people infected. Note that accuracy increases as the number infected increases. The volume, also with a maximum of 0.2 in this case, indicates the accuracy and the proportion of the overall time that is saved by detection through syndromic surveillance as compared to clinical detection. Note the the volume decreases as the number infected increases, mainly because detection through clinical case-finding occurs faster when more people are infected.

and outbreak detection through clinical case-finding in nearly half of the simulated outbreaks.

Perhaps more importantly than interpreting the surface directly, is measuring the volume under the surface (VUS). The VUS is a convenient summary measure for comparing the overall detection accuracy and timeliness of surveillance relative to clinical case-finding.

Table 2 demonstrates that even though the detection accuracy (i.e., AUC) increases as the size of the outbreaks increases, the VUS is lower for larger outbreaks than for small outbreaks. At first this result may seem counterintuitive. One might expect, intuitively, that the accuracy of detection through syndromic surveillance would increase and the time to detection would decrease as the size of the outbreak increases. This is indeed what we observed in the current study when timeliness was measured as the delay from the onset of the outbreak until detection of the outbreak. The VUS, however, measures the timeliness of outbreak detection through syndromic surveillance relative to detection through clinical case-finding. As the size of the outbreak decreased, or the number of infected people decreased, the time until outbreak detection increased for both syndromic surveillance and clinical case-finding, but the increase occured more quickly for clinical case-finding than it did for syndromic surveillance. The net result was that syndromic surveillance saved more time, signalling earlier in the outbreak relative to clinical case-finding, when the outbreak was smaller in size. To some extent, this finding may be influenced by assumptions made in developing our simulation model (see Section 7), but previous work suggests that this finding is not sensitive to modeling assumptions [4, 5].

| Infected | Area Under Curve | Accuracy | Volume Under Surface | Timeliness Accuracy |
|---|---|---|---|---|
| **50% Batch and 50% Real-Time Reporting** | | | | |
| 100 | 0.164 | 0.82 | 0.0934 | 0.47 |
| 500 | 0.166 | 0.83 | 0.0831 | 0.42 |
| 1,000 | 0.173 | 0.86 | 0.0794 | 0.40 |
| 5,000 | 0.195 | 0.97 | 0.0709 | 0.35 |
| **100% Real-Time Reporting** | | | | |
| 100 | 0.156 | 0.78 | 0.0880 | 0.44 |
| 500 | 0.166 | 0.83 | 0.0781 | 0.39 |
| 1,000 | 0.180 | 0.90 | 0.0750 | 0.38 |
| 5,000 | 0.199 | 0.99 | 0.0665 | 0.33 |

TABLE 3. Measures of accuracy and timeliness by outbreak size when using the current mix of batch and real-time reporting as compared to all real-time reporting. Note that timeliness was always better with mixed reporting, while accuracy was higher with real-time reporting at higher numbers infected and lower at lower numbers infected. See the text for further discussion.

3.3. **Detection by Reporting Delay.** An important practical consideration for the surveillance network is whether real-time reporting will result in improved outbreak detection performance over batch reporting of records from emergency departments. Batch reporting is easier to implement in many settings, but if this comes at a cost of worse outbreak detection when compared to real-time reporting, then the additional effort required to implement real-time reporting may be justified.

Our results in this area are mixed, and the reason for these mixed results is not immediately clear (Table 3). Accuracy of detection, as measured by the area under the ROC curve, tended to improve with real-time reporting, but only for larger outbreaks. For smaller outbreaks, the current mix of batch and real-time reporting had greater accuracy than real-time reporting. The results for timeliness, as measured by the volume on the TROC surface, were consistent, with faster detection using the current mix of batch and real-time reporting as compared to all real-time reporting. Further investigation is required to understand the reasons for these findings.

3.4. **Influence of Analysis Frequency.** Another important practical consideration is how frequently the data should be analyzed. The intuition behind more frequent analyses is that this approach may lead to more rapid outbreak detection. The drawback, however, is that analyses performed in rapid succession may provide little new information but they can increase the alarm rate.

In the surveillance network, the data are analyzed 10 times each day. As we noted earlier in this report, multiple analyses each day do not affect the specificity of the system per analysis, but they do tend to increase the alarm rate of the system

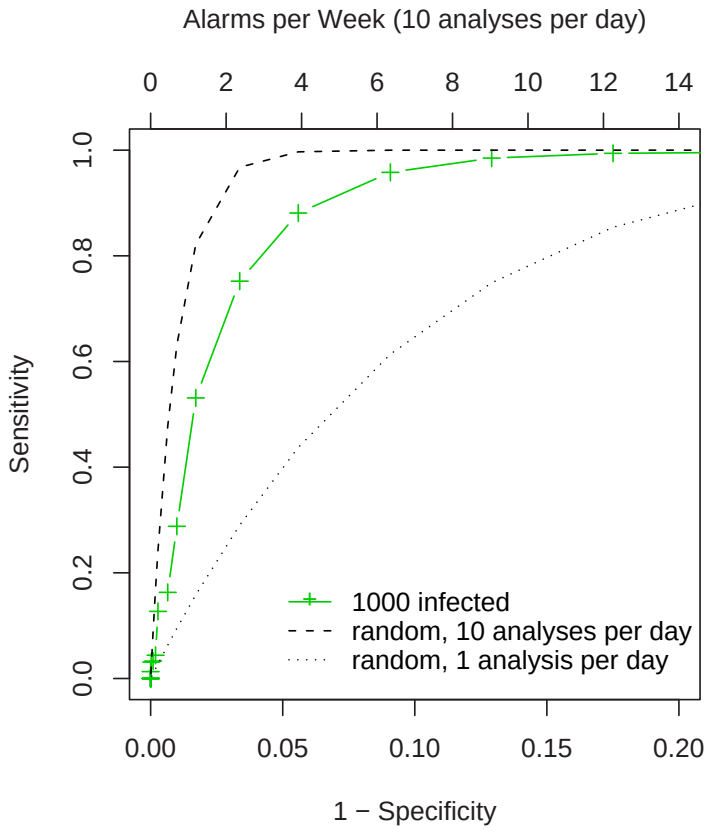


FIGURE 6. The Receiver Operating Characteristics (ROC) Curve for 1,000 people infected and for random alarms at 10 analyses per day and 1 analysis per day. Note that when analyses are conducted ten times each day, a random alarm actually results in better detection performance than the surveillance system. When only one analysis is conducted each day, however, the performance of the random alarm worsens as shown.

per unit of time. As Figure 6 shows, a random analysis, at the same frequency, 10 times each day, has higher accuracy than the current system. When the number of analyses is decreased to once each day, then the performance of random analysis degrades as shown in the Figure. While it is not shown in Figure 6, it is likely that the surveillance system would outperform a random analysis if the data were analyzed once each day.

When plotting the accuracy of any detection method (e.g., a diagnostic test) using a ROC curve, the traditional interpretation is that a random analysis would produce a diagonal curve if both axes of the graph extended from 0 to 1. It is important, therefore, to show how a random analysis results in a different curve

for the surveillance system under evaluation. The random analysis curve in a traditional setting is based on the understanding that 1 statistical test is performed for each true event. In the current study, the outbreaks were on average 10 days in duration, and 10 statistical tests were applied each day, so a total of 100 statistical tests were performed for each true event, or outbreak. Increasing the number of tests performed during each outbreak increases the probability, at a given alarm rate, that a random alarm will detect an outbreak.[5]

A careful analysis to determine the optimal number of analyses each day is beyond the scope of this report. Conducting 1 analysis each day as opposed to 10, however, is likely to decrease timeliness, but the increased time to detection may be outweighed by the decreased alarm rate. We caution against drawing any strong conclusions from these observations about the frequency of analysis in other settings (e.g., for other types of outbreaks), and generalization to other types of analyses (e.g., space-time analysis) is not warranted. Nevertheless, the frequency of routine analysis within the surveillance network is an important topic for future investigation.

## 4. Conclusions and Recommendations

The results from our simulation study suggest that the regional syndromic surveillance reporting network in North Central Texas is capable of detecting a large proportion of inhalational anthrax outbreaks infecting as few as 100 people at false alarm rates acceptable to the surveillance team. For outbreaks of all sizes examined, the surveillance network saved a considerable proportion of time, or detected the outbreak faster, than clinical case-finding for many of the simulated outbreaks. This ability of the surveillance network to provide an early indication of an outbreak was particularly pronounced for smaller outbreaks, where clinical case-finding tended to take longer to identify the sentinel case. In the following sections, we provide specific recommendations for the configuration of the network and for future evaluation.

### 4.1. **Surveillance Network Configuration.**

In general, our results suggest that the surveillance network is currently operating in a manner that is likely to detect disease outbreaks similar to a simulated inhalational anthrax outbreak. We therefore do not suggest any major changes to the current configurations, but there are some minor changes that may streamline operations and improve performance.

First, we noted that the RLS algorithm outperformed the CUSUM algorithm in all inhalational anthrax scenarios examined. In other words, the RLS algorithm always had higher sensitivity at a given alarm rate for a simulated inhalational anthrax outbreak than the CUSUM algorithm and always detected the anthrax outbreaks before the CUSUM algorithm. This finding may be specific to the shape of the epidemic curve for inhalational anthrax and the baseline incidence of respiratory syndromes, but it suggests that greater weight should be given to the results of surveillance alarms using the RLS algorithm for these types of data. For example, when monitoring respiratory syndromes with a high baseline incidence, it would be reasonable for an analyst to conclude that a CUSUM alarm in the absence of an RLS alarm does not provide strong evidence for a true outbreak. In contrast, an

[5]The sensitivity, $Se$, of a random analysis at a set specificity, $Sp$, is calculated from the number of tests performed in an outbreak interval, $nTests$, as $Se = 1 - Sp^{nTests}$.

RLS alarm in the absence of a CUSUM alarm is more likely to indicate a true outbreak. Care should be taken, however, in applying this interpretation to situations where the incidence of cases is low or where the outbreak signal is likely to rise slowly. For example, in detecting sporadic gastrointestinal outbreaks, where the total number of counts may be low relative to the baseline, the CUSUM algorithm may have performance that is superior to the RLS algorithm.

Second, we are not able to say definitively, for the purpose of rapid outbreak detection, whether additional effort should be expended to increase the proportion of hospitals reporting in real time as opposed to batch. There may be other reasons for increasing the proportion of hospitals reporting in real time, such as contributing to 'situational awareness', and we are not commenting on the importance of real-time reporting for purposes other than outbreak detection. We do suggest, however, that the current analysis protocol for batch data in the RODS system be examined carefully to confirm our findings that analysis of batch data may, paradoxically, result in higher accuracy and faster detection than analysis of real-time data in some settings. If our findings are verified, modification to the batch analysis protocol may be warranted.

Third, we suggest that consideration be given to the number of analyses conducted each day. Frequent, routine analyses may increase the chance of false positive alarms and add little additional information. The alarm rates described in this report, while acceptable to surveillance analysts in the network, could be cut in half by decreasing the number of analyses each day from 10 to 5, or decreased by an order of magnitude by decreasing the number of analyses each day to 1. This would not, of course, preclude *ad hoc* inspection of the data as required, it would decrease only the number of routine analyses.

4.2. **Future Evaluation Studies.** The results from our evaluation study suggest a number of areas for future inquiry. Most notably, we examined statistical outbreak detection only, and it is important to understand the implications of our findings in the context of the broader function of the surveillance network. This would include the evaluation of the influence of alarm rate and other factors on protocols for investigating outbreaks and on the use of information from the system to inform public health decision-making.

There are also many aspects of outbreak detection that require additional evaluation. Some possible areas of future study include:

- Conducting follow-on evaluation to address specific issues raised in the current study, including the influence on outbreak detection of: the proportion of all hospitals included in the surveillance network, the proportion of hospitals reporting in real time and in batch mode, and the frequency of analysis.
- Analyzing the performance of other applications used in North Central Texas in addition to RODS (e.g., ESSENCE, RedBat and BioSense).
- Addressing the interplay of different syndromes, space-time analytic methods, and statistical methods that allow concurrent analysis of multiple syndromes (all valuable follow-on work because our study examined temporal surveillance only).
- Examining methods of making the response protocols more efficient and effective. (This might mean, for example, determining whether it's possible to automate some of the steps taken currently by surveillance analysts

to rule out alarms. If such automation is possible, then it may be possible to lower alarm rates while maintaining high levels of sensitivity and timeliness.)

## 5. Author Biographical Sketches

5.1. **David Buckeridge.** M.D., Ph.D., is an Assistant Professor of Epidemiology and Biostatistics at McGill University in Montreal where he holds a Canada Research Chair in Public Health Informatics. Dr. Buckeridge is also a practicing public health physician, working as a medical consultant for the Montreal Public Health Department and the Public Health Institute of Quebec with a mandate to automate surveillance practices. He has a M.D. from Queen's University in Canada, a M.Sc. in Epidemiology from the University of Toronto, and a Ph.D. in Biomedical Informatics from Stanford University. Dr. Buckeridge is also a Fellow of the Royal College of Physicians and Surgeons of Canada with specialty training in Community Medicine.

Dr. Buckeridge has conducted extensive work in the field of evaluating outbreak detection in public surveillance systems. He led an effort to develop an outbreak simulation model for the DARPA-sponsored Bio-ALIRT biosurveillance program and conducted outbreak simulation studies to inform surveillance efforts by the Department of Homeland Security. Dr. Buckeridge's research on outbreak detection in public health surveillance systems has been published in journals such as *Annals of Internal Medicine*, *Statistics in Medicine*, and the *Journal of the American Medical Informatics Association*.

5.2. **Aman Verma.** M.Sc., is a graduate of Computer Science from Queen's University in Canada and also holds a Master of Health Informatics degree from Dalhousie University. Mr. Verma has preformed extensive work in the area of Health Informatics while working in Suriname, South America for the Pan-American Health Organization. He led a team to computerize the Malaria Prevention Program for the country. This project involved extensive work in Geographical Information Systems (GIS) to map data of malaria incidence to identify the best areas of the jungle to send doctors to. In addition, he built an AIDS database to track the status and number of patients within the country. Mr. Verma has also worked with the College of Pharmacy at Dalhousie University in developing a web application to track the health status of diabetes patients. Prior to his work in health informatics, Mr. Verma obtained industrial experience in software development while working with Nortel Networks, where he wrote simulation software to evaluate network components.

5.3. **David Siegrist.** Ph.D., is a Senior Research Fellow at the Potomac Institute for Policy Studies, a not-for-profit research group in Arlington, VA. Dr. Siegrist holds a Ph.D. in Biodefense from George Mason University, a M.Sc. in Management from National Louis University and a M.A. in Government from Georgetown University. Most recently he has been the principal investigator for a project to evaluate how effectively medical biosurveillance can support outbreak detection around the nation in an effort called the BioWatch Signal Interpretation and Integration Project (BWSIIP). Dr. Siegrist has previously led an evaluation with the Bio-ALIRT biosurveillance program in which a large scale data experiment was

performed in order to determine the sensitivity, specificity and timeliness of syndromic surveillance systems to detect naturally-occurring outbreaks of respiratory and gastrointestinal disease. The results were published in a supplement to the CDC journal *Morbidity and Mortality Weekly Report*.

Dr. Siegrist has previously managed and led all aspects of three studies at the Institute on biodefense supported by private foundations and government agencies, and been a leading contributing author. The first volume was an in-depth study on the multi-disciplinary challenge of countering biological terrorism, including analyzing terrorist groups and capabilities to use biological agents. The second study was conducted in coordination with the National Defense University and included sponsorship by the Department of Justice. It identified and prioritized advanced technology needs for biological terrorism consequence management. The third study, *Technologically-Based Biodefense*, identified promising technical approaches to meet those needs.

Dr. Siegrist was a featured speaker and chairman of the threat panel that considered terrorist groups, capabilities and intentions at the First National Symposium on Medical and Health Response to Biological Terrorism in 1998. A written version of his remarks, "Reality of the Threat: Why is there Concern Now?" was published in the Center for Disease Control's journal, *Emerging Infectious Diseases*.

## 6. Acknowledgements

We thank the members of the Tarrant County Public Health Department and the RODS development team for their assistance throughout the project. We also thank Ken Kleinman for providing input on the study protocol and the evaluation methods used in the study.

This publication was produced under the direction of the Southwest Center for Advanced Public Health Practice (Center) at Tarrant County Public Health, and was supported by Cooperative Agreement Number U50/CCU302718 from the Centers for Disease Control and Prevention (CDC) to the National Association of County and City Health Officials (NACCHO). Its contents are solely the responsibility of Center and do not necessarily represent the official views of CDC or NACCHO.

## 7. Appendix - Detailed Methodology

7.1. **Study Design.** To evaluate outbreak detection by the surveillance network, we used the 'inject' approach. This entailed generating many simulated disease outbreaks, superimposing each simulated outbreak onto real baseline data from the surveillance network, and then feeding the combination of real and simulated data to a model of the syndromic surveillance network.

For baseline data, we acquired from the regional syndromic surveillance reporting network in North Central Texas records of visits to emergency departments (ED) between July 2004 and March 2006. Each record contained the date and time of admission to the ED, the name of the ED, and a free-text chief complaint. We used the CoCo chief-complaint classifier, as supplied by the RODS lab along with a training file, to classify each record in the baseline data into a single syndrome.

For the simulated records, we examined eight scenarios defined by the number infected (100, 500, 1,000, and 5,000) and the nature of data acquisition (real-time reporting versus a mix of real-time and batch reporting). For each scenario, we conducted 1,000 simulations, resulting in 4,000 simulated outbreaks in total. We superimposed each outbreak in turn onto baseline data from the year 2005 with the outbreak beginning on a randomly selected date and at a randomly selected time, and we applied two temporal outbreak detection algorithms from the RODS system to the time-series formed by aggregating the baseline and simulated records across the entire surveillance region.

Finally, we calculated the performance of the algorithms and we compared the performance of syndromic surveillance to clinical case-finding through routine blood culture. The main outcomes were the sensitivity, specificity, and timeliness of outbreak detection through syndromic surveillance, and the detection benefit of syndromic surveillance over clinical case-finding.

In the following sections we describe our methods for generating outbreak signals (Section 7.2), our simulation model (Section 7.3), our approach to combining simulated and baseline data (Section 7.4), the outbreak detection algorithm we used (Section 7.5), and the metrics we used to evaluate outbreak detection performance (Section 7.6).

7.2. **Generation of Simulated Signals.** To generate the simulated outbreaks for each scenario, we first specified manually the number infected. We then generated a disease path for each infected individual, the timing of visits to EDs for symptomatic individuals, the syndrome assigned at each visit, and the occurrence, timing and results of blood culture testing. Table 4 shows the parameter values used in the simulation study. The same random number generator with the same seed value was used for each scenario. We used a combined multiple recursive generator as proposed and implemented by L'Ecuyer with the default initial seed [11]. This sampling strategy was intended to improve the efficiency of the simulation and reduce the variance of the output variables [10]. The net result is to facilitate comparison of the results across the different scenarios.

7.3. **Simulation Model.** The simulation model comprises two components: disease and health-care utilization. The first component, **disease**, simulates how infected individuals progress through distinct disease states. The **health-care utilization** component then identifies when symptomatic individuals seek care, their presenting syndromes, and the timing and results of blood-culture testing. In this

| Parameter | Distribution | Parameter Value | Source |
|---|---|---|---|
| **Holding-Time Functions** | | | |
| *Disease Model* | | | |
| Incubation (days), median | lognormal | 10.95 | [3] |
| Incubation, dispersion∗ | lognormal | 2.05 | [3] |
| Prodromal (days), median | lognormal | 2.5 | [7] |
| Prodromal, dispersion | lognormal | 1.44 | [7] |
| Fulminant (days), median | lognormal | 1.5 | [20] |
| Fulminant dispersion | lognormal | 1.44 | [20] |
| *Health-Care Utilization Model* | | | |
| Time until visit | right triangular | disease-state duration† | Estimate |
| Time until blood culture growth (days) | exponential | 0.86 | [1] |
| Time until blood culture isolation (days) | exponential | 1.0 | [8] |
| **Transition Probabilities** | | | |
| *No Visit to Visit ($\alpha_s$)* | | | |
| Prodromal state visit ($\alpha_p$) | Bernoulli | 0.09 | [14] |
| Fulminant state visit ($\alpha_f$) | Bernoulli | 0.80 | Estimate |
| *Visit to Growth ($\beta_s = \beta_s^1 \times \beta^2$)* | | | |
| Blood culture test, prodromal state ($\beta_p^1$) | Bernoulli | 0.1 | [6] |
| Blood culture test, fulminant state ($\beta_f^1$) | Bernoulli | 0.5 | [6] |
| Sensitivity of blood culture ($\beta^2$) | Bernoulli | 0.9 | [16] |
| *Growth to Isolation ($\gamma$)* | | | |
| Prodromal and fulminant states | Bernoulli | 0.8 | [2] |
| **Assignment of Syndrome** | | | |
| Probability of respiratory, prodromal state | Bernoulli | 0.8 | [7, 9] |
| Probability of gastrointestinal, prodromal state | Bernoulli | 0.1 | [7, 9] |
| Probability of neurological, prodromal state | Bernoulli | 0.1 | [7, 9] |
| Probability of respiratory, fulminant state | Bernoulli | 0.7 | [7, 9] |
| Probability of neurological, fulminant state | Bernoulli | 0.3 | [7, 9] |

TABLE 4. Parameter values used in the simulation study. The indicated source was used to identify the best estimate of the parameter value. ∗Following Sartwell [17], the parameter $d = e^s$, where $s^2$ is the variance, is referred to as the *dispersion factor*. †The length of disease state was used to parameterize the triangular visit-time distribution.

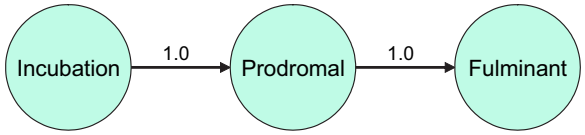


FIGURE 7. The disease model for inhalational anthrax. Infected individuals all pass through three disease states. In the incubation state individuals have no symptoms. In the prodromal state individuals experience an influenza-like illness. Finally, in the fulminant state individuals experience severe symptoms, such as shock. The holding-time function in each state is a lognormal distribution with parameters shown in Table 4.

section, we describe in detail the simulation model components including the values used to parameterize each component for the study.

7.3.1. *Disease.* The disease model takes as input the number of infected individuals and returns a disease path for each individual. The disease path describes the amount of time spent in each discrete disease state. We used a semi-Markov process to model progression of an individual with inhalational anthrax through three disease states: incubation, prodromal, and fulminant (Figure 7) [20].

The definition of the semi-Markov process requires identification of the states, including the holding-time functions, and specification of the transition probabilities between states. The initial state in the model was incubation, followed by certain transition to the prodromal state, and then the fulminant state. For holding-time functions, we used the lognormal distribution, which appears to describe the duration of incubation for many diseases [15, 17], including inhalation anthrax [3, 13]. The values used to parameterize the holding-time functions are taken from observational studies of human exposure [3, 13] and other modeling studies [19, 20], and are shown in Table 4.

7.3.2. *Health-Care Utilization.* The health-care utilization model takes as input a set of disease paths and for each path performs three tasks: (1) it identifies if and when individuals seek care in each disease state, (2) it determines the presenting syndrome for individuals that seek care, and (3) it identifies the timing and results of blood culture testing once care is sought.

We used a semi-Markov process to model health-care utilization (Figure 8). A separate process was used to describe health-care utilization in each of the prodromal and fulminant disease states. Both processes had the same states and transitions (Figure 8), but some values for holding-time functions and transition probabilities differed between the disease states (i.e., those states with a $s$ subscript in Figure 8) and the values used in the simulation study are shown in Table 4.

The transition from 'No Visit' to 'Visit' represents an individual seeking care at an emergency department. The probability of this transition occurring ($\alpha_s$) differs between disease states ($s$). We set the probability of a visit in the prodromal disease state ($\alpha_p$), to 0.09 because cross-sectional surveys suggest that this proportion of individuals visit an ED at some point during an episode of upper respiratory tract illness [12, 14]. For the fulminant disease state ($\alpha_f$), we estimated the probability of seeking care as 80% given the severity of the symptoms in that state.

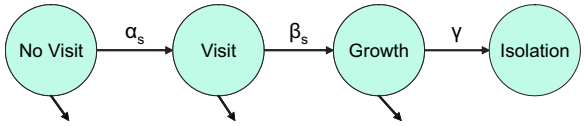


FIGURE 8. The health-care utilization model. When individuals enter the prodromal or fulminant disease state, they enter the health-care utilization model. The probability of making a visit ($\alpha_s$) varies by disease state (Table 4). The probability of a positive blood culture ($\beta_s$) is the product of the probability of ordering a test (which varies by disease state) and the sensitivity of the test (which does not vary by disease state). The probability of isolating the organism does not vary by disease state. The holding-time function for the 'No Visit' state is triangular, with a duration equivalent to the length of the disease state. Holding-time functions for the 'Visit' and 'Growth' states are exponential, with parameters shown in Table 4.

The transition from 'Visit' to 'Growth' represents an individual having a positive blood culture test after making a visit. The probability of this transition ($\beta_s$) is the product of the probability of performing a blood culture test ($\beta_s^1$) and the sensitivity of the test ($\beta^2$, i.e., $\beta_s = \beta_s^1 \times \beta^2$). The probability of performing a test in the prodromal state ($\beta_p^1$), was estimated from the National Ambulatory Health Care Survey as approximately 0.1 [6]. In the fulminant state, we used the same source to estimate the probability of a blood culture test ($\beta_f^1$) as 0.5. We relied on published studies of blood-culture testing to estimate the sensitivity of blood-culture testing in both symptomatic disease states ($\beta^2$) as 0.8 [16].

The final transition, from 'Growth' to 'Isolation', represents the decision to isolate the organism from a blood culture bottle growing gram-positive rods. We relied on data from a recent survey to estimate this value ($\gamma$) as 0.9 [2].

In addition to a transition probability, each of the first three states in the health-care utilization model also requires a holding-time function. The holding-time function for the 'No Visit' state models the distribution of time to seeking care, given that care is sought. We used a right triangular distribution fit to the time spent in the disease state. So, for example, if an individual had a prodromal disease state duration of 10 days, then the probability of seeking care at the instant of entering the disease state would be zero, and the probability of seeking care would increase linearly to 0.2 at ten days, with a mean time to seeking care of 6.7 days.[6] This approach to modeling visits effectively limits individuals to a single visit in each disease state. The selection of a triangular distribution reflects the lack of published evidence about the timing of health-care utilization following the onset of symptoms.

The holding time function for 'Visit' reflects the distribution of times until growth occurs given that the test is positive. The holding time function for 'Growth' is the distribution of times until the organism is isolated given that a decision is made to

[6]These values result from the properties of the triangular distribution, which is defined by three parameters: a, b, and c. In a right triangular distribution, b = c. The maximum point density is 2 / (b - a) and the mean is (a + b + c) / 3.

isolate a specific organism. We modeled both these holding times as exponential with means obtained from published reviews of blood-culture testing [1, 8].

Finally, for individuals that made a health care visit, we simulated the syndrome assigned to the individual using probabilities that reflect the distribution of clinical presentations for inhalational anthrax reported in the literature [7–9]. The probability of being assigned a specific syndrome in the prodromal and fulminant disease states is shown in Table 4.

7.4. **Combination of Simulated Data with Baseline Data.** We first defined a 365 day interval on the baseline data from January 1, 2005 until December 31, 2005 as possible starting dates for a simulated outbreak. We then selected randomly 1,000 dates and times for injecting outbreaks. For simplicity, we will describe these instances of exposure as a 'date', but we sampled a random date and time for each exposure.

To inject the simulated outbreak signals for a scenario, we used the following method. Each scenario had set of simulated outbreaks, $O = \{O_1, \ldots, O_{1000}\}$, and the set of randomly ordered dates, $D = \{D_1, \ldots, D_{1000}\}$. For outbreak $j$, $j \in \{1, \ldots, 1000\}$, we selected $O_j$ and $D_j$. The outbreak $O_j$ is a time series of counts, lasting $n$ days and representing the visits for respiratory, neurological, and gastrointestinal syndromes from the day of the simulated release until the day of the peak incidence of cases.

The outbreaks series $O_j$ is an ordered set of values, $O_j = \{o(1), \ldots, o(n)\}$, which we define to run from $D_j$ until $D_j + n - 1$, or $O_j = \{o(D_j), \ldots, o(D_j + n - 1)\}$. The background time series is also an ordered set of values, $B = \{b(1), \ldots, b(m)\}$. For inject $j$, we extract a subset of the background time series $B_j = \{b(D_j - g), \ldots, b(D_j), \ldots, b(D_j + n - 1)\}$, where $g$ is the length of the lead-in gap, which is the amount of time in the inject series before the beginning of the outbreak. We then define the inject series $I_j = \{i(D_j - g), \ldots, i(D_j + n - 1)\}$, with the entries in the series defined as,

$$i(k) = \begin{cases} b(k) + o(k) & \text{if } k \geq D_j \text{ and } k < D_j + n \\ b(k) & \text{otherwise} \end{cases}$$

In other words, the inject series is formed by adding the values of the outbreak series to the values of the background series, after aligning the two series so that the first day of the outbreak series is added to day $D_j$ in the background series. We then applied the outbreak detection algorithms to each day in the inject series to generate alarm values from day $D_j - g$ to day $D_j + n - 1$. The lead-in gap $g = 180$ days was used, which corresponds to the amount of historical data used by the detection algorithms.

7.5. **Outbreak Detection.** For outbreak detection through syndromic surveillance, we used software implementations of the RLS and Cusum algorithms provided to us by algorithm developers at the RODS laboratory. We were able to adjust the specificity or false alarm rate of each algorithm by altering an alarm threshold parameter. An algorithm was said to have detected an outbreak if the algorithm output a value over its threshold at any point between the onset of exposure and the peak daily incidence of visits during the outbreak (see Section 7.6 for a formal definition of an alarm). The time until outbreak detection for an algorithm was the interval between exposure and the first alarm declared.

As is the current policy in the surveillance network, we applied each algorithm to the combined baseline and outbreak data 10 times each day. The first analysis each day occurred at 6 a.m. and then analyses were conducted once every 2 hours until midnight. To estimate the effect of reporting on outbreak detection, we conducted each analysis using 2 models of record reporting as described in Section 7.5.1.

For outbreak detection through clinical case-finding, we took the minimum time to clinical diagnosis through blood-culture in each simulation run as the time until detection.

7.5.1. *Models of Real-Time and Batch Reporting.* In modeling the arrival of records for analysis by the surveillance network, we used only the time-stamp for the admission of the patient into the emergency department. For real-time reporting, we assumed that records were available instantaneously for analysis at the time of admission. We ignored, therefore, any delay in reporting and any latency in network transmission. For batch reporting, we assumed that records for a 24-hour interval ending at midnight were available instantaneously for analysis at midnight. In practice, however, the first analysis to incorporate batch records was the analysis conducted at 6 a.m.

For each simulated outbreak we conducted 2 sets of analyses. For the first set of analyses, all records were available for analysis in real time. For the second set of analyses, approximately half of the records were available for analysis in real time and the remaining batch-reported records were available for analysis as a group according to the batch reporting protocol described in the previous paragraph. The first data set was intended to reflect the best possible reporting scenario and the second data set was intended to reflect the current reporting situation for the surveillance network.

7.6. **Evaluation of Outbreak Detection Performance.** To measure outbreak detection performance, we calculated, over a range of algorithm thresholds (h) for all surveillance system / detection algorithm pairings, false alarm rate, sensitivity, timeliness, and detection benefit. We provide a precise definition of each metric below, and the definitions rely on the concept of the alarm value $A$ for an algorithm at a given threshold $h$ for each interval analyzed $j$. The alarm value is a binary measure, or

$$A(h)_j = \begin{cases} 1 & \text{if } S(h,j) > h \\ 0 & \text{otherwise.} \end{cases}$$

where $S(h,j)$ is some value returned from the algorithm after analyzing the interval $j$ with threshold $h$.

7.6.1. *Specificity.* Specificity is the probability of no alarm given that there is no outbreak, or

$$Specificity = P(\overline{A}|\overline{O}) = \frac{n(\overline{A},\overline{O})}{n(\overline{O})},$$

where $n(\overline{O})$ is the number of intervals (e.g., days) in the baseline data and $n(\overline{A},\overline{O})$ is the number of alarms when the algorithm is applied to the test data without any superimposed outbreaks. We calculated specificity at a decision threshold $h$ as:

$$Sp(h) = \frac{1}{m}\sum_{j=1}^{m} A(h)_j,$$

where there were $m$ intervals in the baseline data. Note that specificity was calculated using only non-outbreak, or baseline, data. To determine specificity, we applied an algorithm at a given threshold to a time-series of data aggregated across the entire study region. We assumed, therefore, that any alarm in the baseline data was a false alarm in the sense that it did not reflect a true outbreak of inhalational anthrax.

7.6.2. *Sensitivity.* Sensitivity is the probability of an alarm given an outbreak, or

$$Sensitivity = P(A|O) = \frac{n(A,O)}{n(O)},$$

where $n(O)$ is the number of outbreaks and $n(A,O)$ is the number of outbreaks during which an alarm was sounded. We calculate sensitivity at a decision threshold $h$ over some number $n$ of test data sets $i$ as:

$$Se(h) = \frac{1}{n}\sum_{i=1}^{n} \min(1, \sum_{j=1}^{m_i} A(h)_{ij}),$$

where there are $m_i$ intervals in test data set $i$. Note that sensitivity measures only whether an alarm occurred at any point during an outbreak and the timing of an alarm within an outbreak interval is not measured by sensitivity.

7.6.3. *Timeliness.* Timeliness is calculated for a single simulated outbreak as:

$$T(h,i) = \min_j(j : A(h)_{ij} = 1),$$

where there are $m_i$ intervals $i$ and timeliness is not defined if $\sum_{j=1}^{m_i} A(h)_{ij} = 0$.

7.6.4. *Detection Benefit.* Detection benefit is the potential gain in time to detection from using one detection method relative to another method. For one detection method $A$ and another method $B$, the benefit of $A$ over $B$ is calculated as the difference in the timeliness using the two methods, or

$$D_{AB}(h,i) = \max(0, T_B(h,i) - T_A(h,i)).$$

The detection benefit is always greater than or equal to zero. If method $B$ always detects an outbreak, which is the case in the current study for clinical case-finding, then the detection benefit is the difference in timeliness between the two methods when method $A$ detects an outbreak.

Another useful metric is the proportion of time saved [**?**], and an alarm saving at least $tsp$ of the proportion of time is defined as,

$$A(h, tsp) = A(h) \bullet I(D_{AB} \leq tsp),$$

and the true positive rate is defined as

$$TP(h, tsp) = \sum_{n=1}^{N} \frac{A_n(h, tsp)}{N},$$

where $A_n(h, tsp)$ is an alarm for outbreak $n$ at threshold $h$ that saves at least the proportion $tsp$ of time and there are $N$ outbreaks.

## References


1. SE Beekmann, DJ Diekema, KC Chapin, and GV Doern, *Effects of rapid detection of bloodstream infections on length of hospitalization and hospital charges.*, Journal of Clinical Microbiology **41** (2003), no. 7, 3119–25.
2. EM Begier, *Gram-positive Rod Surveillance for Early Anthrax Detection.*, Emerg Infect Dis **11** (2005), no. 9, 1483–1486.
3. R Brookmeyer, N Blades, M Hugh-Jones, and DA Henderson, *The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak*, Biostatistics **2** (2001), no. 2, 233–247.
4. DL Buckeridge, *A method for evaluating outbreak detection in public health surveillance systems that use adminstrative data*, Ph.D. thesis, Stanford University, 2005.
5. DL Buckeridge, P Switzer, D Owens, D Siegrist, J Pavlin, and M Musen, *An evaluation model for syndromic surveillance: assessing the performance of a temporal algorithm*, MMWR Morb Mortal Wkly Rep **54 Suppl** (2005), 109–15, 1545-861X (Electronic) Journal Article.
6. National Center for Health Statistics, *Public use data tapes: National ambulatory medical care survey 2003*, Tech. report, National Center for Health Statistics, 2003.
7. JE Holty, DM Bravata, H Liu, RA Olshen, KM McDonald, and DK Owens, *A century of inhalational anthrax: A systematic review of cases from 1900 to 2005*, Annals of Internal Medicine **To Appear** (2006).
8. TV Inglesby, T O'Toole, DA Henderson, JG Bartlett, MS Ascher, E Eitzen, AM Friedlander, J Gerberding, J Hauer, J Hughes, J McDade, MT Osterholm, G Parker, T M Perl, P K Russell, and K Tonat, *Anthrax as a biological weapon, 2002: updated recommendations for management.*, Journal of the American Medical Association **287** (2002), no. 17, 2236–52.
9. JA Jernigan, DS Stephens, DA Ashford, C Omenaca, MS Topiel, M Galbraith, M Tapper, TL Fisk, S Zaki, T Popovic, RF Meyer, CP Quinn, SA Harper, SK Fridkin, JJ Sejvar, CW Shepard, M McConnell, J Guarner, WJ Shieh, JM Malecki, JL Gerberding, JM Hughes, BA Perkins, and Anthrax Bioterrorism Investigation Team, *Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States*, Emerging Infectious Diseases **7** (2001), no. 6, 933–44.
10. AM Law and WD Kelton, *Simulation modeling and analysis*, Third ed., McGraw-Hill, 2000.
11. P L'Ecuyer, *Random number generation*, Handbook of Computational Statistics (JE Gentle, E Härdle, and Y Mori, eds.), Springer-Verlag, 2004, pp. 35–70.
12. WJ McIsaac, N Levine, and V Goel, *Visits by adults to family physicians for the common cold.*, The Journal of Family Practice **47** (1998), no. 5, 366–9.
13. M Meselson, J Guillemin, M Hugh-Jones, A Langmuir, I Popova, A Shelokov, and O Yampolskaya, *The Sverdlovsk anthrax outbreak of 1979*, Science **266** (1994), no. 5188, 1202–8.
14. KB Metzger, A Hajat, M Crawford, and F Mostashari, *How many illnesses does one emergency department visit represent? Using a population-based telephone survey to estimate the syndromic multiplier.*, MMWR Morbidity and Mortality Weekly Report **53 Suppl** (2004), 106–11.
15. P Philippe, *Sartwell's incubation period model revisited in the light of dynamic modeling.*, Journal of Clinical Epidemiology **47** (1994), no. 4, 419–33.
16. LG Reimer, ML Wilson, and MP Weinstein, *Update on detection of bacteremia and fungemia.*, Clinical Microbiology Reviews **10** (1997), no. 3, 444–65.
17. PE Sartwell, *The distribution of incubation periods of infectious diseases*, American Journal of Hygiene **51** (1950), 310–318.
18. FC Tsui, JU Espino, VM Dato, PH Gesteland, J Hutman, and MM Wagner, *Technical description of RODS: a real-time public health surveillance system*, Journal of the American Medical Informatics Association **10** (2003), no. 5, 399–408.
19. GF Webb and MJ Blaser, *Mailborne transmission of anthrax: Modeling and implications*, Proceedings of the National Academy of Sciences USA **99** (2002), no. 10, 7027–32.
20. LM Wein, DL Craft, and EH Kaplan, *Emergency response to an anthrax attack*, Proceedings of the National Academy of Sciences USA **100** (2003), no. 7, 4346–4351.